# **MAT 303 Project Two Summary Report**

Justin Farquhar

Justin.farquhar@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## **1. Introduction**

For this project we will be exploring a data set containing risk factors for heart disease. This larger data set will allow us to analyze patterns between these different health indicators and the presence of heart disease. These results can help the hospital better analyze each patient’s risk and if further actions need to be taken to help prevent heart disease. This will help boost patient care as well as quality of life for these patients. We will be utilizing two logistic regression models as well as a random forest classification model and a random forest regression model.

## **2. Data Preparation**

There are 303 rows and 14 variables/columns in the data set. The data set contained many more variables but the important ones that we will be utilizing through this study are age, sex, cp or type of chest pain experienced, trestbps or resting blood pressure, chol or cholesterol, fbs or fasting blood sugar, restecg or resting electrocardiographic measurement, thalach or maximum heart rate achieved, exang or exercise-induced angina, oldpeak or ST depression induced by exercise, slope or the slope of the peak exercise ST segment, ca or the number of major vessels, and target which is the presence of heart disease.

## **3. Model #1 - First Logistic Regression Model**

### **Reporting Results**

General Form

Prediction Equation

Linear Form

Within the linear form we have the presence of the π, which represents the probability of the event occurring. We also have 1- π, which represents the probability of the event not occurring. The instance of function could realistically be replaced with odds in the linear form that was provided, but I wanted to leave them in to reference their meaning here. To then move onto interpreting the estimated coefficient of maximum heart rate achieved, the model gives us the value of 0.031095 which means that for every unit increase of maximum heart rate achieved, the odds of the person having heart disease increases by 0.031095.

### **Evaluating Model Significance**

The null hypothesis for a logistic regression model is always going to be that the model fits the data well and the alternative hypothesis will be that the model does not fit the data well. The P value that was returned was 0.612, which is higher than the 5% level of significance, indicating that there is insufficient evidence to reject the null hypothesis, meaning that the model fits the data well. The test also returned a test statistic of 44.622 which is slightly lower than the Chi-square distribution of 48 degrees, which shows that there is no significant deviation from the observed and predicted value. Looking through each individual predictors significance level, we have a value of 0.0107 for age, a value of -0.0012 for trestbps, a value of -1.1220 for exang1, and a value of 0.0431 for thalach. This shows that every variable is significant at the 5% level of significance other than exang1 which ended up being extremely negative.

For the model, we returned an accuracy value of 0.7360, a precision value of 0.7322, and a recall value of 0.8121.

A graph of a function

Description automatically generated

The above image is the ROC curve of the model. Generally, when we would see the curve going quickly up to the left corner, the model would suggest a high sensitivity and specificity which would mean that it distinguishes between the classes well. However, in our case, this is not true. The curve is in the positive direction, but it begins heading to the right too soon. Through this we also got an area under the curve value of 0.8007. In the case of a perfect classifier, this value would be 1 while a random classifier would have a value of 0.5. With our value being 0.8007, it is fairly average, leaning more towards being good than being truly random.

### **Making Predictions Using Model**

*A white background with black text

Description automatically generated*

The image above represents the individual predictions that we made with the model. For the first prediction of an individual having heart disease who is 50 years old, has a resting blood pressure of 122, has exercise induced angina, and has a maximum heart rate of 140 we get a 27.16% likelihood of them having heart disease. For the second prediction of an individual having heart disease who is 50 years old, has a resting blood pressure of 130, does not have an exercise induced angina, and has a maximum heart rate of 165 we get a 78.53% likelihood of them having heart disease. These results actually had me doubting the validity of the model all together. Through my previous experience, an individual with exercise induced angina and a lower maximum heart rate is going to be at a much higher risk than someone without exercised induced angina and a higher maximum heart rate. For maximum heart rate, it slightly depends on the circumstance of the measurement. If this was obtained during exercise, someone able to have a higher heart rate is generally seen as healthier. However, a resting heart rate that is higher is not seen as healthy.

## **4. Model #2 - Second Logistic Regression Model**

### **Reporting Results**

General Form

Prediction Equation

Linear Form

### **Evaluating Model Significance**

The null hypothesis for a logistic regression model is always going to be that the model fits the data well and the alternative hypothesis will be that the model does not fit the data well. The P value that was returned was 0.7485, which is higher than the 5% level of significance, indicating that there is insufficient evidence to reject the null hypothesis, meaning that the model fits the data well. The test also returned a test statistic of 41.121 which is slightly lower than the Chi-square distribution of 48 degrees, which shows that there is no significant deviation from the observed and predicted value. Looking through each individual predictors significance level, we have a value of 0.0162 for age, a value of -0.0047 for trestbps, a value of 2.7223 for cp1, a value of 2.6666 for cp2, a value of 2.6785 for cp3, and a value of 0.0438 for thalach. This shows that every variable is significant at the 5% level of significance other than the different chest pain variables.

For the model, we returned an accuracy value of 0.7756, a precision value of 0.8012, and a recall value of 0.7818.

A graph of a function

Description automatically generated

The above image is the ROC curve of the model. Generally, when we would see the curve going quickly up to the left corner, the model would suggest a high sensitivity and specificity which would mean that it distinguishes between the classes well. However, in our case, this is not true. The curve is in the positive direction, but it begins heading to the right too soon. Through this we also got an area under the curve value of 0.8389. In the case of a perfect classifier, this value would be 1 while a random classifier would have a value of 0.5. With our value being 0.8389, it is fairly average, leaning more towards being good than being truly random. This ROC curve is also better than the previous model, as well as having a slightly higher AUC level.

### **Making Predictions Using Model**

*A white background with black text

Description automatically generated*

To then make predictions using the new model, for an individual having heart disease who is 50 years old, has a resting blood pressure of 115, does not experience chest pain, and has maximum heart rate of 133, the model returned a 27.56% chance. Then for an individual having heart disease who is 50 years old, has a resting blood pressure of 125, experiences typical angina, and has a maximum heart rate of 155, the model returned a 82.18% chance. This seems much more in-line with how it should be based on general health data and seems much more accurate than the previous model.

## **5. Random Forest Classification Model**

### **Reporting Results**

For this model we will be moving to a random forest classification model. There were 303 rows within the original data set, there are 257 rows in the training set, and there are 46 rows in the testing set.

A graph of a number of trees

Description automatically generated

Looking at the above graph for the training and testing error against the number of trees, we can see that the line roughly starts to flatten just after 20 trees. In this case, I opted to call it 25.

### **Evaluating the Utility of the model**

To then look at the utility of the model, we can assess the confusion matrix. For the training set, the model returned an accuracy value of 0.9922, a precision value of 0.9927, and a recall value of 0.9927. For the testing set, the model returned an accuracy value of 0.6522, a precision value of 0.7308, and a recall value of 0.6786.

## **6. Random Forest Regression Model**

### **Reporting Results**

To then move onto the random forest regression model we will split the training and testing sets using a 80% and 20% split. This gives 242 rows for the training set and 61 rows for the testing set.

A graph of a number of trees

Description automatically generated

The image above is the mean squared error against the number of trees for our random forest regression model. The line is already fairly flat, but I believe it gets much more consistent around 20 trees, so I opted to use that value for the next tests.

### **Evaluating the Utility of the Random Forest Regression Model**

Using 20 trees for the regression tree the root mean squared error for the training set came out to 11.6282 and the root mean squared error for the testing set came out to 21.1139. This means that with each value, the model was off by a mean of 11.6282 and 21.1139, which is fairly accurate when considering maximum heart rates are usually over 150.

## **7. Conclusion**

I would chose the second logistic regression model to predict heart disease. From my personal experience, the first model was giving the opposite result of what I would expect for the predictions, while the second model had a much better result. The scripts for the random forest classification model are much more complicated than the logistic models, but I would absolutely recommend them over the logistic models. The accuracy, precision, and recall values were each much higher, indicating that they are much better predictors of heart disease. For the practical importance of the analyses, being able to better predict the odds of heart disease can help determine if a patient needs to be fully screened for heart disease. As for some cases, the severity of heart disease is unknown until it is far too late. This can be extremely detrimental for a patient and could even lead to their death. With these results these situations can be wildly avoided through proper screenings. These studies could also open the door for future analyses for other diseases that could help avoid other unnecessary harm to patients.